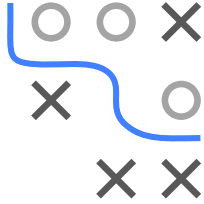
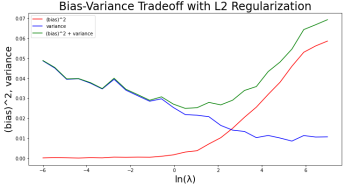
Introduction to Machine Learning



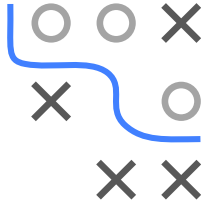
Regularization
Perspectives on Ridge Regression
(Deep-Dive)



Learning goals

- Know interpretation of *L2* regularization as row-augmentation
- Know interpretation of *L2* regularization as minimizing risk under feature noise
- Bias-variance tradeoff for ridge regression

PERSPECTIVES ON *L2* REGULARIZATION



We already saw two interpretations of *L2* regularization.

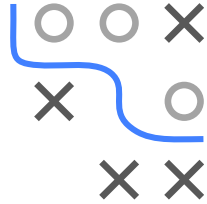
- We know that it is equivalent to a constrained optimization problem:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{ridge}} &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)}\right)^2 + \lambda \|\boldsymbol{\theta}\|_2^2 = (\mathbf{X}^T\mathbf{X} + \lambda \boldsymbol{I})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 \text{ s.t. } \|\boldsymbol{\theta}\|_2^2 \leq t\end{aligned}$$

- Bayesian interpretation of ridge regression: For normal likelihood contributions $\mathcal{N}(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}, \sigma^2)$ and i.i.d. normal priors $\theta_j \sim \mathcal{N}(0, \tau^2)$, the resulting MAP estimate is $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ with $\lambda = \frac{\sigma^2}{\tau^2}$:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\theta} \log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta})] = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)}\right)^2 + \frac{\sigma^2}{\tau^2} \|\boldsymbol{\theta}\|_2^2$$

L2 AND ROW-AUGMENTATION

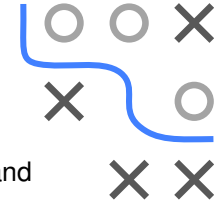


We can also recover the ridge estimator by performing least-squares on a **row-augmented** data set: Let $\tilde{\mathbf{X}} := \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\boldsymbol{I}_p \end{pmatrix}$ and $\tilde{\mathbf{y}} := \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_p \end{pmatrix}$. Using the augmented data, the unregularized least-squares solution $\tilde{\boldsymbol{\theta}}$ can be written as

$$\begin{aligned}
\tilde{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n+p} \left(\tilde{y^{(i)}} - \boldsymbol{\theta}^T \tilde{\mathbf{x}^{(i)}} \right)^2 \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 + \sum_{j=1}^{p} \left(0 - \sqrt{\lambda}\theta_j \right)^2 \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 + \lambda \|\boldsymbol{\theta}\|_2^2
\end{aligned}$$

$\Longrightarrow \hat{\boldsymbol{\theta}}_{\text{ridge}}$ is the least-squares solution $\tilde{\boldsymbol{\theta}}$ but using $\tilde{\mathbf{X}}, \tilde{\mathbf{y}}$ instead of $\mathbf{X}, \mathbf{y}$!

L2 AND NOISY FEATURES



Now consider perturbed features $\tilde{\mathbf{x}^{(i)}} := \mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)}$ where $\boldsymbol{\delta}^{(i)} \overset{iid}{\sim} (\mathbf{0}, \lambda \boldsymbol{I}_p)$. Note that no parametric family is assumed. We want to minimize the expected squared error taken w.r.t. the perturbations $\boldsymbol{\delta}$:

$$\mathcal{R}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\delta}}\Big[\frac{1}{n}\sum_{i=1}^{n}\big((y^{(i)} - \boldsymbol{\theta}^\top \tilde{\mathbf{x}^{(i)}})^2\big)\Big] = \mathbb{E}_{\boldsymbol{\delta}}\Big[\frac{1}{n}\sum_{i=1}^{n}\big((y^{(i)} - \boldsymbol{\theta}^\top (\mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)}))^2\big)\Big] \;\Big|\; \text{expand}$$

$$\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\delta}}\Big[\frac{1}{n}\sum_{i=1}^{n}\big((y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 - 2\boldsymbol{\theta}^\top \boldsymbol{\delta}^{(i)}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}) + \boldsymbol{\theta}^\top \boldsymbol{\delta}^{(i)} \boldsymbol{\delta}^{(i)\top} \boldsymbol{\theta}\big)\Big]$$

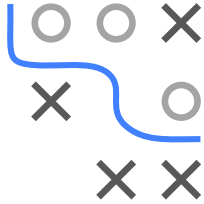
By linearity of expectation, $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}^{(i)}] = \mathbf{0}_p$ and $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}^{(i)} \boldsymbol{\delta}^{(i)\top}] = \lambda \boldsymbol{I}_p$, this is

$$\begin{aligned}\mathcal{R}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\big((y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 - 2\boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}^{(i)}](y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)}) + \boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}^{(i)} \boldsymbol{\delta}^{(i)\top}]\boldsymbol{\theta}\big) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \lambda \|\boldsymbol{\theta}\|_2^2\end{aligned}$$

$\Longrightarrow$ Ridge regression on unperturbed features $\mathbf{x}^{(i)}$ turns out to be minimizing squared loss averaged over feature noise distribution!

BIAS-VARIANCE DECOMPOSITION FOR RIDGE

For linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ with $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, bias of ridge estimator $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ is given by

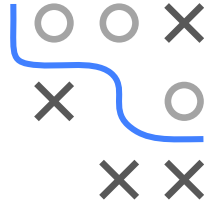


$$
\begin{aligned}
\text{Bias}(\hat{\boldsymbol{\theta}}_{\text{ridge}}) := \mathbb{E}[\hat{\boldsymbol{\theta}}_{\text{ridge}} - \boldsymbol{\theta}] &= \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}] - \boldsymbol{\theta} \\
&= \mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon})] - \boldsymbol{\theta} \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} + (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \underbrace{\mathbb{E}[\boldsymbol{\varepsilon}]}_{=0} - \boldsymbol{\theta} \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{\theta} \\
&= \left[(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right] \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta}
\end{aligned}
$$

- Last expression shows bias of ridge estimator only vanishes for $\lambda = 0$, which is simply (unbiased) OLS solution
- It follows $\|\text{Bias}(\hat{\boldsymbol{\theta}}_{\text{ridge}})\|_2^2 > 0$ for all $\lambda > 0$

BIAS-VARIANCE DECOMPOSITION FOR RIDGE / 2

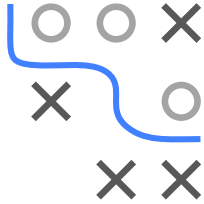
For the variance of $\hat{\boldsymbol{\theta}}_{\text{ridge}}$, we have



$$
\begin{aligned}
\text{Var}(\hat{\boldsymbol{\theta}}_{\text{ridge}}) &= \text{Var}\left((\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right) \quad \Big| \text{ apply } \text{Var}_u(\boldsymbol{A}\boldsymbol{u}) = \boldsymbol{A}\text{Var}(\boldsymbol{u})\boldsymbol{A}^\top \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \text{Var}(\boldsymbol{y}) \left((\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top\right)^\top \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \text{Var}(\varepsilon) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \sigma^2 \boldsymbol{I}_n \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\
&= \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}
\end{aligned}
$$

- $\text{Var}(\hat{\boldsymbol{\theta}}_{\text{ridge}})$ is strictly smaller than $\text{Var}(\hat{\boldsymbol{\theta}}_{\text{OLS}}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ for any $\lambda > 0$, meaning matrix of their difference $\text{Var}(\hat{\boldsymbol{\theta}}_{\text{OLS}}) - \text{Var}(\hat{\boldsymbol{\theta}}_{\text{ridge}})$ is positive definite (bit tedious derivation)
- This further means $\text{trace}\big(\text{Var}(\hat{\boldsymbol{\theta}}_{\text{OLS}}) - \text{Var}(\hat{\boldsymbol{\theta}}_{\text{ridge}})\big) > 0 \, \forall \lambda > 0$

BIAS-VARIANCE DECOMPOSITION FOR RIDGE / 3



With bias and variance of the ridge estimator we can decompose its mean squared error as follows:

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\text{ridge}}) = \|\mathrm{Bias}(\hat{\boldsymbol{\theta}}_{\text{ridge}})\|_2^2 + \mathrm{trace}\big(\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\text{ridge}})\big)$$

Comparing MSEs of $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ and $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ and using $\mathrm{Bias}(\hat{\boldsymbol{\theta}}_{\text{OLS}}) = 0$ we find

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\text{OLS}}) - \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\text{ridge}}) = \underbrace{\mathrm{trace}\big(\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\text{OLS}}) - \mathrm{Var}(\hat{\boldsymbol{\theta}}_{\text{ridge}})\big)}_{>0} - \underbrace{\|\mathrm{Bias}(\hat{\boldsymbol{\theta}}_{\text{ridge}})\|_2^2}_{>0}$$

Since both terms are positive, sign of their diff is *a priori* undetermined.
Theobald 1974 and Farebrother 1976 prove there always exists some $\lambda^* > 0$ so that

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\text{OLS}}) - \mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\text{ridge}}) > 0$$

Important theoretical result: While Gauss-Markov guarantuees $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ is best linear unbiased estimator (BLUE), there are biased estimators with lower MSE.